DOCUMENT RESUME

ED 442 821		TM 031 241

AUTHOR	Walker, Cindy M.; Beretvas, S. Natasha
TITLE	Using Multidimensional versus Unidimensional Ability Estimates To Determine Student Proficiency in Mathematics.
PUB DATE	2000-04-00
NOTE	34p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Elementary School Students; Intermediate Grades; *Junior High School Students; Junior High Schools; *Mathematics Achievement; Mathematics Tests; State Programs; Test Items; Testing Programs
IDENTIFIERS	*Item Dimensionality; *Mathematical Communication; Multidimensionality (Tests); Unidimensionality (Tests)

ABSTRACT

The primary objective of this research was to examine the effect of scoring items known to be multidimensional using a unidimensional model. Although several simulation studies have examined this, few studies have been conducted using data obtained from actual test administrations. In this study, open-ended mathematics items from a mandated state test, previously shown to function differentially in favor of proficient writers, were hypothesized to be multidimensional. (Data were obtained from 65,333 fourth graders and 65,279 seventh graders taking the state mathematics tests.) Only these items comprised the second dimension, considered to be mathematical communication, while all of the mathematics items defined both the unidimensional model and the first factor of the multidimensional model, considered to be general mathematical ability. The pattern of examinee placement into four different proficiency level classifications, previously determined using the bookmark standard setting procedure, was compared for both the unidimensional model and the first dimension of the multidimensional model. The majority of examinees placed into different levels was placed into higher levels of proficiency by the multidimensional model. Further analyses indicated that the average level of mathematical communication differed for examinees placed into different levels by the two models. Examinees with higher estimates of mathematical communication tended to be placed into a higher proficiency level, while those with lower estimates of mathematical communication tended to be placed into lower proficiency levels by the unidimensional model. (Contains 8 tables and 16 references.) (Author/SLD)

Running Head: Using Multidimensional Ability Estimates

Using Multidimensional versus Unidimensional Ability Estimates

to Determine Student Proficiency in Mathematics

Cindy M. Walker

S. Natasha Beretvas

University of Washington

This paper was presented at the 2000 Annual Meeting of the American Educational Research Association, New Orleans, LA. Correspondence should be addressed to: Cindy M. Walker, University of Washington, College of Education, 312 Miller, Box 353600, Seattle WA 98195-3600, (206) 616-6305, cwalker@u.washington.edu

Abstract

The primary objective of this research was to examine the effect of scoring items known to be multidimensional using a unidimensional model. Although several simulation studies have examined this, few studies have been conducted using data obtained from actual test administrations. In this study, open-ended mathematics items from a mandated state test, previously shown to function differentially in favor of proficient writers, were hypothesized to be multidimensional. Only these items comprised the second dimension, considered to be mathematical communication, while all of the mathematics items defined both the unidimensional model and the first factor of the multidimensional model, considered to be general mathematical ability. The pattern of examinee placement into four different proficiency level classifications, previously determined using the bookmark standard setting procedure, was compared for both the unidimensional model and the first dimension of the multidimensional model. The majority of examinees placed into different levels were placed into higher levels of proficiency by the unidimensional model. Further analyses indicated that the average level of mathematical communication differed for examinees placed into different levels by the two models. Examinees with higher estimates of mathematical communication tended to be placed into a higher proficiency levels, while those with lower estimates of mathematical communication tended to be placed into lower proficiency levels by the unidimensional model.

Using Multidimensional versus Unidimensional Ability Estimates to Determine Student Proficiency in Mathematics

The primary purpose of any mathematics assessment is to quantify the unobservable construct commonly referred to as "*mathematical ability*". As with any psychological measurement endeavor, no single approach to quantifying this construct is universally agreed upon. Rather many different approaches exist, such as performance assessments, portfolios, journals, observational assessments, and multiple choice standardized tests. While few would argue that the alternatives to standardized multiple choice tests should play an important part in any mathematics classroom, the use of these alternative assessments for large-scale testing seems daunting due to the time and money needed for such an undertaking. Yet multiple choice standardized tests in mathematics have traditionally been criticized for their perceived misalignment with the curriculum, their inability to provide information about the process of student learning, as well as the widespread belief that the use of these tests can negatively affect the quality of mathematics instruction (Romberg, Zarinnia, & Collins, 1990; Shepard, 1992; Silver & Kenney, 1995).

Despite the criticisms of multiple choice standardized tests, the typical format of most items used for large-scale assessment purposes, the public still clamors for the results of such tests, the results of which are routinely published in newspapers. This is most likely due to the fact that the public educational system affects each and every one of us in some way. State legislators use the results as a measure of public schools for accountability purposes. School principals may use the results as a measure of teachers for accountability purposes. Researchers commonly use the results as evidence of students' learning. Parents may use the results when deciding where to purchase a home. However, only to the extent that the test is aligned with what we think our students should be learning in school can we make any valid inferences

regarding what actually is being *taught* in the schools using these results.

In the realm of mathematics it seems clear what students should be learning. The publication of *The Curriculum and Evaluation Standards* (National Council of Teachers of Mathematics [N.C.T.M.], 1989), *The Professional Standards for Teaching Mathematics* (N.C.T.M., 1991) and *The Assessment Standards for School Mathematics* (N.C.T.M., 1995) have provided educators with a vision of what it means to know and understand mathematics. These documents, commonly referred to as the *NCTM Standards*, emphasize the need for mathematics students to spend more time on mathematical reasoning and problem solving, communicating mathematical ideas, exploring relationships among representations of mathematical forms, and making connections between mathematical topics.

Many testing programs have addressed the earlier criticisms of standardized mathematics tests in two ways. First, most large-scale mathematics tests now include more items devoted to problem solving, reasoning, and non-standard mathematical topics. Ironically, now some of these tests have been criticized for their lack of symbolic computational problems. Second, although limited by available resources, some large-scale testing programs have also incorporated open-ended polytomous items, in addition to the traditional multiple-choice format, to try and capture the *process* of student learning, in addition to the *product.*

The state of Washington could be considered a leading example in this movement. In creating their assessment system for mathematics they began by first communicating to the public what students should know and be able to do in the subject of mathematics. These descriptions, known as the Essential Academic Learning Requirements [EALRs] (Commission on Student Learning, 1995) closely parallel the vision outlined in the *NCTM Standards* (NCTM,

1989, 1991, 1995). Only after the criteria, or curriculum standards, were made public did the state embark on creating a criterion-referenced test based on the requirements set forth in the EALRs.

However, these innovations in testing result in a myriad of possible psychometric problems. For example, traditionally standard item response theory has been used for scaling these tests. One of the assumptions of item response theory is that of unidimensionality. It would seem that emphasis on problem solving and mathematical reasoning and the intentional inclusion of mathematical communication would result in a multidimensional test. This is due to the fact that in order for a student to do well on items that are measuring problem solving and reasoning, students must also be able to read and comprehend the problems. Moreover, in order for students to do well on items measuring mathematical communication students must also be able to communicate clearly graphically, numerically, and in writing. Nevertheless, unidimensional models are still being used for practical scaling purposes, even though multidimensional models exist.

When a unidimensional model is used to fit multidimensional data, several problems can arise. Simulation studies conducted by Way, Ansley, & Forsyth (1988) have shown that the estimate of the discrimination parameter obtained by fitting a two parameter logistic (2-PL) model, when the data comes from a non-compensatory multidimensional model, is comparable to the *average* of the two discrimination parameters that would have been obtained if a non-compensatory model were fit. Moreover, it was found that if the data comes from a compensatory multidimensional model the discrimination parameter obtained by fitting a 2-PL model is comparable to the *sum* of the two discrimination parameters that would have been found if a compensatory model were fit. Furthermore, the difficulty parameters obtained by

fitting a 2-PL model tend to overestimate the true difficulty parameter when a multidimensional model is called for (Way, Ansley, & Forsyth, 1988).

Problems can also arise when estimating an examinee's level of ability. If a unidimensional model is used when a multidimensional model is more appropriate, an examinee's unidimensional estimate of ability is actually a linear combination of the ability estimates that would be obtained if a multidimensional IRT model were used (Ackerman, 1994). Furthermore, if difficulty and dimensionality are confounded in the data, this composite of ability does not remain consistent throughout the estimated unidimensional ability scale (Reckase, Carlson, Ackerman, & Spray, 1986). To make matters even worse, when groups of examinees differ in their underlying distributions on these traits, yet only a single score is reported, differential item functioning (DIF) occurs (Ackerman & Evans, 1994). This implies that if items function differentially then the construct being measured is multidimensional. One interpretation of this implication suggests that mathematics items that measure problem solving, requiring examinees to read and interpret a problem situation, should function differentially in favor of examinees who are better able to comprehend what they have read. This paradigm also suggests that mathematics items that require examinees to communicate about mathematics in writing should function differentially in favor of examinees who are better able to organize their ideas in writing.

This second hypothesis was substantiated using data from the 1998 administration of the Washington Assessment of Student Learning (WASL) administered to fourth and seventh graders (Walker & Beretvas, 1999). Students were grouped based on writing proficiency using only those items from the writing section of the test that were designed to measure organizational skills in writing. Two groups were formed that contrasted highly capable students with those

who were extremely non-proficient. Open-ended mathematics items that required students to communicate about mathematics were chosen *a priori* and bundled together to conduct the DIF analyses, as suggested by Roussos and Stout (1996), with the thought that these items would function differentially in favor of students who were highly capable of organizing and expressing their ideas in writing. For both grade levels the results strongly supported this hypothesis.

Theoretically, these results suggest that two scores should be reported for the mathematics items shown to be multidimensional: one representing an examinee's ability to communicate about mathematics and another representing an examinee's ability to solve mathematical problems. However, currently only one score is reported. What is the effect of using only this one score to make inferences regarding an examinee's ability in mathematics? Would using two scores result in different student rankings and/or different diagnostic conclusions? Perhaps using only a single score results in a student being labeled as not meeting the standard in mathematics, when in reality this student is just not meeting the standard in mathematical communication. This research addresses these questions. The primary objective of this research was to examine the effect of scoring items known to be multidimensional in a unidimensional manner.

Methods

Participants

This research utilized data obtained from fourth and seventh graders who participated in the 1998 administration of the WASL. Only students who were not given any type of accommodation (i.e. mainstream students) were considered. This resulted in 65, 333 eligible fourth grade examinees and 65, 279 eligible seventh grade examinees. All of the eligible examinees were considered in the final analyses, however approximately 30,000 from each

group were randomly sampled to use in item calibration for each test.

Instrumentation

The mathematics component of the WASL was used as the measure of mathematical ability. For fourth graders, this particular form of the test consisted of 24 multiple-choice items and 16 open-ended items. For seventh graders, this particular form of the test consisted of 30 multiple-choice items and 16 open-ended items. All open-ended items were hypothesized to be multidimensional because they required students to communicate about mathematics and were previously shown to function differentially in favor of proficient writers (Walker & Beretvas, 1999). These items required students to either 1) explain their thinking using words, numbers or pictures; 2) describe a graph or table or use this information to write mathematical problems; or 3) explain the logic presented in a problem that may or may not be correct.

Methodology

The scores on the open-ended mathematics items were dichotomized. A review of the original scoring rubrics associated with four-point extended response mathematics items indicated that a score of four was assigned to a response that met all relevant criteria, while a score of three was assigned to a response that met all or most relevant criteria. Moreover, a response that only met some or few relevant criteria and may have omitted information was assigned a score of two and one respectively. A score of zero was assigned to a response that showed no understanding of the problem. (Office of the Superintendent of Public Instruction [OSPI], 1999). For these four-point extended response items, scores of zero, one and two were re-coded to a value of zero, while scores of three and four were re-coded to a value of one. Only four of the seventh grade items and three of the fourth grade items were scored using this five category schema.

The remaining thirteen open-ended items for the fourth grade examinees and twelve open-ended items for the seventh grade examinees were scored using a three-category schema. For these items a score of two was assigned to responses that showed clear understanding and complete analysis or interpretation. A score of one was assigned to responses that were incomplete or ineffective and showed only partial understanding. A score of zero was assigned to responses that demonstrated little or no understanding, including such responses as "I don't know or "?" (OSPI, 1999). For these items, scores of zero and one were re-coded to a score of zero, while scores of two were re-coded to a score of one.

NOHARM II (Fraser & McDonald, 1988) was used to fit both a unidimensional and a multidimensional normal ogive model. Guessing parameters need to be fixed and input by the user. To obtain the guessing parameters for the multiple-choice items MULTILOG VI (Thissen, 1991) was used to calibrate these items. The guessing parameters for the open-ended items were assumed to be zero. For the unidimensional case this is comparable to fitting the following three parameter logistic model (Hambleton & Swaminathan, 1985):

$$P_i(X=1\mid\theta)=c_i+(1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

where:

$P_i(X=1\mid\theta)$ = the probability that an examinee with estimated ability θ answers item i correctly

c_i = the guessing parameter of item i

b_i = the difficulty parameter of item i

a_i = the discrimination parameter of item i

D = the scaling factor of 1.7

For the multidimensional case this is comparable to fitting the following compensatory model (McKinley & Reckase, 1983):

$$P_i\left(X=1 \mid \vec{\theta}\right)=\frac{e^{\left(d_i+\vec{a}_i\vec{\theta}\right)}}{1+e^{\left(d_i+\vec{a}_i\vec{\theta}\right)}}$$

where:

$P_i\left(X=1 \mid \vec{\theta}\right)$ = the probability that an examinee with an estimated vector of abilities, $\vec{\theta}$, obtains a correct answer to item i

$\vec{a}_i$ = a row vector of discrimination parameters for item i

$$d_i = -\sum_{k=1}^{m} a_{ik} b_{ik}$$

where a_{ik} = the discrimination parameter for item i on dimension k

b_{ik} = the difficulty parameter for item i on dimension k

m = the number of dimensions

Due to the evidence suggesting DIF for the open-ended mathematics items (Walker & Beretvas, 1999), confirmatory analyses were conducted when fitting the multidimensional model. Two dimensions were assumed, with all of the items loading on the first dimension and only the open-ended items loading on the second dimension. The first dimension can be interpreted as general mathematical ability while the second dimension can be interpreted as a more specific aspect of mathematical ability, mathematical communication. Since the underlying factors were both measuring different aspects of mathematical ability, the two underlying factors were assumed to be correlated.

A standard setting procedure, comparable to the bookmark standard setting procedure (Lewis et. al., 2000), was conducted by the state of Washington to define the minimum number correct score for an examinee to be considered proficient in mathematics. Four different levels,

two levels of non-proficiency and two levels of proficiency, were defined in this process and it is these four levels that are reported (OSPI, 1999). The cut-off for level 3 corresponds to the cut-off for whether or not an examinee is considered to be meeting the standard in mathematics. This bookmark standard setting procedure involves forming a committee of experts who first take the exam as it is administered to examinees (Lewis et. al., 2000). The state of Washington used committees composed of teachers, curriculum specialists, school administrators, parents and community members at large (OSPI, 1999). After the committee has taken the exam they are then given the items re-ordered based on their level of difficulty so that the easiest items appear first. Polytomous items appear more than once. A polytomous item with k categories, defined as 0, 1, 2, ... k will appear k - 1 times in the ordered list. For example, an item with three categories, 0, 1,and 2, will appear once to determine the location of a score of 1 and once to determine the location of a score of 2 (Lewis, et. al., 2000). The state of Washington used Rasch modeling to determine the difficulty of items, although it is conceivable that other models could have been used. Members of the committee are then asked to establish the minimum level of competency students must demonstrate to be categorized at each level. This is done by asking committee members to come to a consensus about the location of the item, in the ordered list of items, for which a student at each level would answer all the preceding items correct with 2/3 probability of success (Lewis, et. al., 2000, OSPI, 1999).

The results of the standard setting analyses conducted by the state of Washington were used to determine the minimum level of estimated ability needed for each level. Specifically, these results were used to find the ability estimates associated with the minimum number correct score that needed to be obtained by an examinee to be categorized into levels 2, 3, and 4. For the unidimensional model only one estimate was found for each of the cut points, $\hat{\theta}_k$, where k = 2,

3, or 4, and refers to the corresponding level. For the multidimensional model two estimates were found for each of the cut points, $\hat{\theta}_{jk}$, where j = 1 or 2 and refers to the dimension and, as before, k = 2, 3, or 4 and refers to the corresponding level. For both models these estimates of ability were obtained by finding the maximum of the likelihood function, which, for the multidimensional case is expressed by:

$$L(u_1, u_2, \cdots, u_n \mid \theta_1, \theta_2) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}$$

where:

u_i = 0 if an item was answered correctly and 1 if an item was answered incorrectly

P_i = the probability of obtaining the correct answer to item i

$Q_i = (1 - P_i)$ = the probability of not answering item i correctly.

The response vector corresponding to the number correct score associated with the cut point for each proficiency level was used in the above equation. Although different possible response vectors correspond to the same number correct score, the vector corresponding to easiest combination of point values was used to find the cut points in this study. For example, the fourth grade mathematics examination contained 24 multiple-choice items, 13 three-category items and 3 five-category items. Therefore, the highest possible number correct score that could be obtained for this was 62. For this grade level, the standard setting committee determined that if an examinee obtained a score of 28 on the mathematics items then they should be assigned level 2 proficiency. Likewise a score of 38 was determined to correspond to level 3 proficiency (i.e. meeting the standard) and a score of 47 was determined to correspond to level 4 proficiency.

Table 1 depicts the response vectors associated with each of these number correct scores that were used to generate the ability estimate cutoffs associated with levels 2, 3, and 4.

Insert Table 1 About Here

The response vectors in Table 1 contain point values for only those items that were determined to be the easiest (i.e. lower difficulty parameters obtained in item calibration). The first three vectors in Table 1 contain both polytomous and dichotomous items because these are the actual items the judges considered. The first 24 responses correspond to dichotomous items, and the last 16 responses correspond to polytomous items with responses in the 28^{th}, 30^{th}, and 39^{th} position associated with five-category items and the other responses in the last 16 positions associated with three-category items. These vectors were dichotomized, in the same way that the data was, to obtain the minimum ability estimates corresponding to the cut points for each of the proficiency levels in the dichotomized data. These dichotomized vectors are also presented in Table 1. These vectors were used to estimate $\hat{\theta}_k$ for the unidimensional model as well as $\hat{\theta}_{jk}$ for the multidimensional model.

For the dichotomized data the highest number correct score a fourth grade examinee could obtain was 40. For this grade level, dichotomizing the polytomous vectors resulted in an observed score of 15 associated with the cut point for level 2, an observed score of 22 associated with the cut point for level 3 (i.e. meeting the standard), and an observed score of 29 associated with level 4 proficiency. The difference between minimum number correct scores for each of the levels in the polytomous case is 10 points, whereas for the dichotomous case the difference

between minimum number correct score is only 7 points. The cut points for seventh grade examinees were found in a similar manner.

For the unidimensional case, students were then assigned to level 4 proficiency if their estimated ability was greater than or equal to $\hat{\theta}_4$. Similarly, students were assigned to level 3 proficiency if their estimated ability was greater than or equal to $\hat{\theta}_3$ and were assigned to level 2 proficiency if their estimated ability was greater than or equal to $\hat{\theta}_2$. For the multidimensional case, there were two cut points that needed to be considered for each level *k*, $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$, one for each dimension. Several different classification schemes exist for this situation. Students could have been considered to be at level *k* if *both* of their ability estimates for the two dimensions were greater than or equal to the ability cut points, $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$. However, with this approach information is lost since in reality a student could be meeting the standard on one dimension but not the other. Alternatively, students could have been considered to be at level *k* if the weighted average of their ability estimates was greater than or equal to some pre-defined weighted average of the ability cut points based on substantive reasoning. However, even with substantive reasoning, the comparative weighted average seems somewhat arbitrary. Another possible classification scheme would be to consider each of the dimensions separately. In other words, a student could be categorized into level *k* on the dimension *j* if their ability estimate for dimension *j* was greater than or equal to the ability cut point, $\hat{\theta}_{jk}$. This approach provides additional diagnostic information, pertaining to the construct of mathematical communication, and therefore was the approach considered in this research.

Results

In order to determine the effect of dichotomizing the data, proficiency levels were determined using the polytomous data, assuming the three parameter logistic model for the multiple-choice items and the generalized partial credit model for the open-ended items. For this classification scheme, the estimated ability cut points were obtained using the polytomous vectors presented in Table 1. These classifications were then compared to the classifications obtained when fitting the unidimensional model to the dichotomized data and using the estimated ability cut points obtained using the dichotomous vectors presented in Table 1. Although some mismatches were found, the majority of examinees were placed in the same proficiency levels for both models. For fourth grade examinees there was 78% agreement between the proficiency level a student would be placed in using the polytomous classification scheme and the level a student would be placed in using the dichotomous classification scheme. Similarly for seventh grade examinees there was 81% agreement.

Table 2 displays the estimated ability cut points for both the unidimensional and the multidimensional models for both fourth and seventh grade levels. Recall that for the multidimensional model it was assumed that the two underlying factors were correlated. For the fourth grade examinees this estimated correlation is 0.61, while for the seventh grade examinees $r = 0.81$. Within both grade levels the estimated ability cut points are quite similar for the unidimensional model and the first dimension of the multidimensional model. For the fourth grade examinees the estimated ability cut points for the second dimension, representing mathematical communication, at proficiency levels 2 and 3 are not that distinct. Similarly for the seventh grade examinees the estimated ability cut points for the second dimension for proficiency levels 3 and 4 are not that dissimilar.

Insert Table 2 About Here

Recall that the first dimension of the multidimensional model can be thought of as representing general mathematical ability while the second dimension can be thought of as representing mathematical communication. The unidimensional model, on the other hand, can be thought of as representing some linear combination of general mathematical ability and the sub-skill of mathematical communication, although this composite of ability may not remain constant throughout the estimated unidimensional ability scale. Table 3 shows the results of classifying fourth grade examinees into the four different proficiency levels based on either the unidimensional model or the first dimension of the multidimensional model. As the table illustrates, the majority of mismatched examinees were placed into lower levels when the first dimension of the multidimensional model was used, as opposed to the unidimensional model. 20.89 % of fourth grade examinees placed into level 3 proficiency under the unidimensional model were placed into level 2 proficiency based on the first dimension of the multidimensional model. A similar pattern is found for fourth grade examinees who were placed into levels 2 and 4 proficiency based on the unidimensional model. A smaller percentage of examinees would have been placed into higher levels of proficiency based on the first dimension of the multidimensional model. 2.39% of examinees classified as level 2 under the unidimensional model and 6.44% of examinees classified as level 3 under the unidimensional model would have been placed into levels 3 and 4, respectively.

Insert Table 3 About Here

Table 4 shows the corresponding results for seventh grade examinees. For this grade level only 47.33% of the examinees placed into level 2 proficiency when the unidimensional model is used were placed into the same level of proficiency based on the first dimension of the multidimensional model. Once again, the majority of the mismatched examinees are placed into lower proficiency levels when the first dimension of the multidimensional model is used. Of those examinees placed into level 2 when the unidimensional model was used, 40.81% and 11.86% were placed into levels 1 and 3, respectively, based on the first factor of the multidimensional model. Similarly, a more examinees placed into level 3 proficiency when the unidimensional model is used would be placed into a lower level of proficiency based on the first dimension of the multidimensional model. 7.23% of these examinees were placed into level 1 and 11.15% of these examinees were placed into level 2, while only 3.92 % of these examinees were placed into level 4 based on the multidimensional model. Interestingly, 5.52% of examinees classified as level 4 proficiency when the unidimensional model is used would be classified as level 1proficiency, although none of these examinees would go down to level 2 proficiency, if the first factor of the multidimensional model were used. 14% of these examinees go down to level 3 proficiency under the multidimensional model.

Insert Table 4 About Here

To further explore what was causing examinees to be classified into different proficiency levels when different models were used, examinees of both grade levels were placed into one of three different groups depending on the cross-classification tables for the two models. The first group was comprised of examinees that were placed into a higher proficiency level when the unidimensional model was used, relative to the proficiency classification based on the first

dimension of the multidimensional model. 7,214 of the fourth grade examinees and 8,361 of the seventh grade examinees were placed into this group. The second group was comprised of examinees that were placed into the same group based on the two different models. 54,208 of the fourth grade examinees and 54,639 of the seventh grade examinees fell into this group. The third group was comprised of examinees that were placed in a lower proficiency group based on the unidimensional model. 2,111 of the fourth grade examinees and 2,279 of the seventh grade examinees were placed into this group.

It was hypothesized that these groups differed in their distributions on the second dimension of the multidimensional model, mathematical communication. Specifically, it was thought that the reason examinees were placed into a higher proficiency level when the unidimensional model was used was because these examinees had a higher level of mathematical communication ability, on average, than other examinees and that the multidimensional model accounted for this ability through the distinct second dimension. Likewise, it was thought that the reason examinees were placed in a lower proficiency level based on the unidimensional model was because these examinees had a lower level of mathematical communication, on average, than other examinees. Finally, it was thought that the reason examinees were placed into the same proficiency level based on the two models was because their level of ability in mathematical communication was similar to the overall average.

To test this hypothesis a one-way ANOVA test were conducted using the estimates of mathematical communication ability as the dependent variable. For both grade levels the results strongly supported the hypothesis. For fourth grade examinees, 11.7% of the variation among the estimates of mathematical communication was explained by between group variation ($F_{2, 63,530} = 4,202.08$, $p < 0.0001$). The estimated effect size for fourth grade examinees was 0.36.

For the seventh grade examinees, 9.1 % of the variation among the estimates of mathematical communication was explained by between group variation $F_{2, 65,276} = 3,256.35$, $p < 0.0001$). The estimated effect size for seventh grade examinees was 0.32. Table 5 displays the means and standard deviations of the estimated mathematical communication level for each of the three groups. To determine which means were significantly different, all pair-wise comparisons were conducted using Tukey's honestly significant difference tests. Table 6 shows the confidence intervals obtained from each of the post-hoc comparisons, each of which was statistically significant. As the table demonstrates, those examinees placed into a lower proficiency level based on the unidimensional model tended to have a lower level of mathematical communication ability, while those placed into a higher level of proficiency based on the unidimensional model tended to have a higher level of mathematical communication ability.

Insert Tables 5-6 About Here

Table 7 displays the cross-tabulation of the classification of fourth grade examinees on both of the dimensions of the multidimensional model: general mathematical ability and mathematical communication. Due to the similarity of the estimated ability cut points for levels 2 and 3 of the second dimension of the multidimensional model for fourth grade examinees only 5.9% of examinees were classified at level 2 proficiency for mathematical communication. Examinees at each level of proficiency on general mathematical ability are found at each level of proficiency on mathematical communication. Furthermore, while only 36.84% of examinees were classified as meeting the standard (i.e. level 3 or level 4) on the dimension of general mathematical ability 73.74% of fourth grade examinees were classified as meeting the standard

on the dimension of mathematical communication.

Insert Table 7 About Here

Table 8 illustrates the same cross-tabulation results for seventh grade examinees. Similar to the results for fourth grade examinees, only 3.75% of seventh grade examinees were placed into level 3 proficiency, presumably because of the similarity of the estimated ability cut points for levels 3 and 4 of the second dimension of the multidimensional model. Similar to the results observed for fourth grade examinees, seventh grade examinees at each level of proficiency on general mathematical ability are found at each level of proficiency on mathematical communication. However fewer seventh grade examinees were found to be proficient in mathematical communication than fourth grade examinees. Only 33.08% of seventh grade examinees were found to be meeting the standard in mathematical communication compared to 73.74% of the fourth grade examinees.

Insert Table 8 About Here

Discussion

The approach taken within this research is not flawless. The fact that the data had to be dichotomized in order to conduct the analyses is a direct result of available models, software, and current estimation procedures, as well as computer limitations. Although multidimensional models for polytomous data based on the Rasch model exist (see Kelderman & Rijkes, 1994 ;

van der Linden & Hambleton, 1997) it is questionable whether ability estimates obtained from Rasch modeling are able to correctly classify examinees since they lack a discrimination parameter. Research has shown that proficiency classifications based on ability estimates obtained from Rasch modeling tend to overestimate ability at the low end of the scale while underestimating ability at the high end of the scale (Beretvas & Walker, 2000). This is primarily due to the fact that there is a one-to-one correspondence between number-correct score and ability estimate when the Rasch model is used. However, when discrimination is accounted for the same number-correct scores can lead to the different ability estimates. Specifically, obtaining the correct answer to less discriminating items will lead to lower ability estimates than obtaining correct answers to items with higher levels of discrimination.

However, the purpose of this research was to compare the use of a unidimensional model on data known to be multidimensional. Previous research had shown both the original polytomous version of the data and the transformed dichotomized data to be multidimensional. Therefore, although the data was transformed it was only the transformed data that was compared under the two models to try and discover the implications of using a unidimensional model when it may not be appropriate. Although much research has been conducted on simulated data, very little research has been done on data taken from real testing situations and given the current limitations the approach taken in this research is one way to explore multidimensional IRT with current tests in use.

Modeling the data in a multidimensional manner allows one to make separate inferences about an examinee for each of the distinct dimensions. This additional information is a valuable asset to anyone who wants to learn more about *why* students are not proficient. This research provides further evidence that when data known to be multidimensional is modeled using a

unidimensional model incorrect inferences may be made about student proficiency. Furthermore, these incorrect inferences are made primarily for those examinees that differ in their ability distributions on the secondary dimension. Specifically, it is these examinees that are more likely to be placed into different proficiency classifications by the different models. Those examinees who tend to have lower estimates of ability on the second dimension of the multidimensional model tend to have lower estimates of ability under the unidimensional model than they would have based on the first dimension of the multidimensional model. Those examinees who tend to have higher estimates of ability on the second dimension of the multidimensional model tend to have higher estimates of ability based on the unidimensional model than they would have based on the first dimension of the multidimensional model. This is true despite the fact that the first dimension of the multidimensional model uses information from the same items that were used for the unidimensional model. These results also provide further evidence to support the multidimensional paradigm for DIF since the items that were chosen to comprise the second dimension exhibited differential item functioning in a previous study.

The fact that the estimated ability cut points were so similar for levels 2 and 3 for the fourth grade examinees is probably why a higher percentage of fourth grade examinees were found to be proficient in mathematical communication. The cut point for level 3 made it extremely easy for a fourth grade examinee to be classified as level 3 proficiency in mathematical communication because it was extremely close to the cut point for level 2 and low in value. Similarly, for the seventh grade examinees the cut points for levels 3 and 4 were extremely close and high in value making it extremely difficult for a seventh grade examinee to be classified as level 3 proficiency, which is the cut point for meeting the standard. There simply

was not a large enough range of ability that led to classification into level 2 for the fourth grade examinees or into level 3 for the seventh grade examinees. This is probably partly due to the dichotomization process. This process led to only a small point difference, on the transformed open-ended items, between levels 2 and 3 for the fourth grade examinees and between levels 3 and 4 for the seventh grade examinees. However, there was also not a very large point difference in the original polytomous open-ended items for these proficiency levels and grade levels. This finding may have implications for standard setting procedures currently in use. If standard setting committee members were asked to set standards based on distinct dimensions a more pronounced difference between each of the cut points would be expected.

In any testing situation for which decisions are to be made on the basis of one test score there are bound to be some misclassifications. No one piece of evidence alone can paint a perfect picture of what a student has learned. There is always some degree of measurement error involved. However, by continuing to model mathematical proficiency using a model that assumes the construct is unidimensional, when we have substantive and empirical reasons to believe mathematical proficiency is a multidimensional construct, we are, perhaps unwittingly, increasing our error of measurement.

REFERENCES

Ackerman, T. A. & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analysis. *Applied Psychological Measurement, 18*(4), 329-342.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18*(3), 257-275.

Beretvas, S. N. & Walker, C. M. (2000). *Proficiency Level Classification Compared Using Rasch versus non-Rasch Models.* Paper presented at the annual meeting of the American Educational Research Association. University of Washington at Seattle.

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika (59)*2, 149-176.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (2000). *The bookmark standard setting procedure: Methodology and recent implementations.* Manuscript submitted for publication.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (1991). *Professional Standards for Teaching Mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (1995). *Assessment Standards for School Mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: Author.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986). *The interpretation of unidimensional IRT parameters estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Canada.

Roussos, L. A. & Stout, W. (1996). A multidimensionality-based DIF paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership 46*, 4-9.

Silver E. A., & Kenney, P. A. (1995). Sources of assessment information in instructional guidance in mathematics. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 38-86). Albany, NY: State University of New York Press.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag New York Inc.

Walker, C. M. & Beretvas, S. N. (1999). An empirical investigation supporting the

multidimensional DIF paradigm: A Cognitive explanation for the occurrence of DIF. Manuscript submitted for publication. University of Washington at Seattle.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 293-252.

Table 1

Response Vectors Used to Estimate Fourth Grade Ability Cut Points for Levels 2 through 4

Level	Original Polytomous Response Vectors
2	1 1 1 1 1 0 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 0 0 1 0 2 1 1 2 1 1 2 2 1 0 0 0 1 2
3	1 1 1 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 2 0 2 1 1 3 1 1 2 2 1 0 1 1 2 2
4	1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1 1 0 1 1 0 1 2 0 2 3 1 3 2 2 2 2 2 0 1 1 4 2
	Revised Dichotomous Response Vectors
2	1 1 1 1 1 0 1 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 1
3	1 1 1 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 1
4	1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 1 1 1 0 0 0 1 1

Table 2

Fourth and Seventh Grade Unidimensional and Multidimensional Estimated Ability Cut Points

	Estimated Ability Cut Points								
	Unidimensional Model			Multidimensional Model					
Grade	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$	$\hat{\theta}_{14}$	$\hat{\theta}_{22}$	$\hat{\theta}_{23}$	$\hat{\theta}_{24}$
Fourth	-0.435	0.229	1.121	-0.337	0.394	1.141	-0.527	-0.427	0.450
Seventh	0.128	0.563	1.197	0.260	0.563	1.270	-0.421	0.333	0.413

Table 3

Cross- Classification of Proficiency Level Placements Under the Unidimensional Model and the First Dimension of the Multidimensional Model for Fourth Grade Examinees (n = 63,533)

		Proficiency Level Classification Based on First Dimension of Multidimensional Model			
		Level 1	**Level 2**	**Level 3**	**Level 4**
Proficiency Level Classification Based on Unidimensional Model	**Level 1**	19,497	505	0	0
		(97.5%)	(2.5%)	(0.00%)	(0.00%)
	Level 2	2,681	13,524	397	0
		(16.15%)	(81.46%)	(2.39%)	(0.00%)
	Level 3	0	3,924	13,648	1,209
		(0.00%)	(20.89%)	(72.67%)	(6.44%)
	Level 4	0	0	609	7,539
		(0.00%)	(0.00%)	(7.47%)	(92.43%)

Note. Reported percentages are row percentages and represent the percentage of examinees who were placed at the proficiency level represented by the row that were placed at each level of proficiency represented by the column.

Table 4

Cross- Classification of Proficiency Level Placements Under the Unidimensional Model and the First Dimension of the Multidimensional Model for Seventh Grade Examinees (n = 65,279)

		Proficiency Level Classification Based on First Dimension of Multidimensional Model			
		Level 1	**Level 2**	**Level 3**	**Level 4**
Proficiency Level Classification Based on Unidimensional Model	**Level 1**	33,603	449	0	0
		(98.68%)	(1.32%)	(0.00%)	(0.00%)
	Level 2	4,614	5,351	1,341	0
		(40.81%)	(47.33%)	(11.86%)	(0.00%)
	Level 3	902	1,392	13,648	489
		(7.23%)	(11.15%)	(77.7%)	(3.92%)
	Level 4	411	0	1,042	5,988
		(5.52%)	(0.00%)	(14.0%)	(80.47%)

Note. Reported percentages are row percentages and represent the percentage of examinees who were placed at the proficiency level represented by the row that were placed at each level of proficiency represented by the column.

Table 5

Mean Mathematical Communication Ability Estimates for Examinees

Fourth Grade Examinees (Overall mean = -0.02, standard deviation = 0.57)			
Group	**n**	**Mean**	**Standard Deviation**
Examinees placed into lower proficiency level based on unidimensional model	7,214	0.46	0.45
Examinees placed into same proficiency levels based on both models	54,208	-0.07	0.55
Examinees placed into higher proficiency level based on unidimensional model	2,111	-0.55	0.45
Seventh Grade Examinees (Overall mean = 0.02, standard deviation = 0.68)			
Group	**n**	**Mean**	**Standard Deviation**
Examinees placed into lower proficiency level based on unidimensional model	8,361	0.58	0.70
Examinees placed into same proficiency levels based on both models	54,639	-0.03	0.65
Examinees placed into higher proficiency level based on unidimensional model	2,279	-0.47	0.41

Table 6

Confidence Intervals for Tukey's h.s.d.Post Hoc Tests for Both Grade Levels

	Fourth Grade Examinees		**Seventh Grade Examinees**	
Group Comparison*	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**
Group 1 – Group 2	0.51	0.53	0.53	0.55
Group 1 – Group 3	0.98	1.01	0.96	0.99
Group 2 – Group 3	0.45	0.48	0.41	0.45

*Note. Group 1 represents examinees that were placed into lower proficiency level based on the unidimensional model. Group 2 represents examinees that were placed into the same proficiency levels based on both models. Group 3 represents examinees that were placed into higher proficiency level based on the unidimensional model.

Table 7

Cross- Classification of Proficiency Level Placements Under the Two Dimensions of the Multidimensional Model for Fourth Grade Examinees (n = 63,533)

		Proficiency Level Classification Based on Second Dimension of Multidimensional Model – Mathematical Communication			
		Level 1	**Level 2**	**Level 3**	**Level 4**
Proficiency Level Classification Based on First Dimension of Multidimensional Model – General Mathematical Ability	**Level 1**	7,768 (35.03%)	2,051 (9.25%)	11,008 (49.63%)	1,351 (6.09%)
	Level 2	3,213 (17.9%)	979 (5.45%)	10,286 (57.29%)	3,475 (19.36%)
	Level 3	1,526 (10.41%)	548 (3.74%)	7,911 (53.99%)	4,669 (31.86%)
	Level 4	410 (4.69%)	185 (2.11%)	4,126 (47.17%)	4,027 (46.03%)

Note. Reported percentages are row percentages and represent the percentage of examinees who were placed at the proficiency level of general mathematical ability represented by the row that were placed at each level of mathematical communication when the multidimensional model is used.

Table 8

Cross- Classification of Proficiency Level Placements Under the Two Dimensions of the Multidimensional Model for Seventh Grade Examinees (n = 65,279)

		Proficiency Level Classification Based on Second Dimension of Multidimensional Model – Mathematical Communication			
		Level 1	**Level 2**	**Level 3**	**Level 4**
Proficiency Level Classification Based on First Dimension of Multidimensional Model – General Mathematical Ability	**Level 1**	17,388 (43.99%)	15,231 (38.53%)	1,144 (2.89%)	5,767 (14.59%)
	Level 2	1,185 (16.48%)	2,907 (40.42%)	419 (5.83%)	2,681 (37.28%)
	Level 3	1,066 (8.82%)	4,148 (34.34%)	547 (4.53%)	6,319 (52.31%)
	Level 4	410 (3.71%)	185 (23.5%)	4,126 (5.2%)	4,027 (67.59%)

Note. Reported percentages are row percentages and represent the percentage of examinees who were placed at the proficiency level of general mathematical ability represented by the row that were placed at each level of mathematical communication when the multidimensional model is used.

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031241

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using Multidimensional versus Unidimensional Ability Estimates to Determine Student Proficiency in Mathematics	
Author(s): Cindy M. Walker, S. Natasha Beretvas	
Corporate Source:	Publication Date: April 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 [X]	Level 2A [ ]	Level 2B [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Cindy M Walker	Printed Name/Position/Title: Cindy M. Walker / Asst. Professor	
Organization/Address: University of Washington	Telephone: 206-616-6305	FAX:
	E-Mail Address: cwalker@u.washington.edu	Date: MAY 4, 2000

(over)

ERIC® **Clearinghouse on Assessment and Evaluation**

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 2000

Dear AERA Presenter,

Congratulations on being a presenter at AERA. The ERIC Clearinghouse on Assessment and Evaluation would like you to contribute to ERIC by providing us with a written copy of your presentation. Submitting your paper to ERIC ensures a wider audience by making it available to members of the education community who could not attend your session or this year's conference.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed, electronic, and internet versions of *RIE*. The paper will be available **full-text, on demand through the ERIC Document Reproduction Service** and through the microfiche collections housed at libraries around the world.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse and you will be notified if your paper meets ERIC's criteria. Documents are reviewed for contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at **http://ericae.net.**

To disseminate your work through ERIC, you need to sign the reproduction release form on the back of this letter and include it with **two** copies of your paper. You can drop of the copies of your paper and reproduction release form at the ERIC booth (223) or mail to our attention at the address below. **If you have not submitted your 1999 Conference paper please send today or drop it off at the booth with a Reproduction Release Form.** Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
The University of Maryland
1129 Shriver Lab
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

ERIC/AE is a project of the Department of Measurement, Statistics and Evaluation at the College of Education, University of Maryland.